Non-response in panel data

Nijman, T.E.; Verbeek, M.J.C.M.

Published in:
Journal of Applied Econometrics

Publication date:
1992

JOURNAL OF APPLIED ECONOMETRICS, VOL. 7, 243–257 (1992)

NONRESPONSE IN PANEL DATA: THE IMPACT ON ESTIMATES OF A LIFE CYCLE CONSUMPTION FUNCTION

THEO NIJMAN AND MARNO VERBEEK
Tilburg University, Department of Economics, PO Box 90153, 5000 LE Tilburg, The Netherlands

SUMMARY

If missing observations in a panel data set are not missing at random, many widely applied estimators may be inconsistent. In this paper we examine empirically several ways to reveal the nature and severity of the selectivity problem due to nonresponse, as well as a number of methods to estimate the resulting models. Using a life cycle consumption function and data from the Expenditure Index Panel from the Netherlands, we discuss simple procedures that can be used to assess whether observations are missing at random, and we consider more complicated estimation procedures that can be used to obtain consistent or efficient estimates in case of selectivity of attrition bias. Finally, some attention is paid to the differences in identification, consistency, and efficiency between inferences from a single wave of the panel, a balanced sub-panel, and an unbalanced panel.

1. INTRODUCTION

One of the almost unavoidable problems in the empirical analysis of panel data is attrition. Individuals initially participating in the panel may drop out after a few waves, or may not be willing or able to participate in some wave, for example because of a holiday. In addition, often new individuals are sampled after a few waves to 'replace' the ones who have dropped out, so as to retain the original sample size as much as possible. The consequence of this is that virtually all available panel data sets are unbalanced.

It is common practice in applied economic analysis of panel data to use only the observations on units for which a complete time-series is available. Since the seminal contributions of Heckman (1976, 1979) and Hausman and Wise (1979) it is well known that the use of complete observations only, can easily yield misleading results originating in inconsistent parameter estimates due to selection bias or attrition bias. In this paper we analyse this problem and illustrate its implications for the analysis of a simple life cycle consumption model. We discuss simple procedures to assess whether several widely applied estimators are consistent, as well as more complicated estimation procedures which can be used to obtain consistent parameter estimates if selectivity due to non-random attrition or nonresponse occurs in the panel. Moreover, some attention is paid to the choice problem of analysing either the complete observations in one wave of the panel only, a balanced panel, containing only those individuals for which a complete time-series is available, or an unbalanced panel.

The plan of this paper is as follows. In Section 2 we introduce the model under consideration, and Section 3 discusses the problem of nonresponse and suggests three

0883–7252/92/030243–15$12.50

Received June 1990
Revised October 1991

possibilities to test for attrition bias. Empirical results based on these procedures are given in Sections 4 to 6. Section 7 contains estimation results which correct for potential attrition bias. Section 8 concludes.

2. MODELLING CONSUMPTION AND NONRESPONSE

In this paper we focus on the effects of nonresponse on estimates of the relationship between total consumption of households and demographic characteristics, such as age and education of the head of the household, the number of adults and the number of children in the household. The model we analyse is:

$$\log C_{it} = \beta_0 + x_{it}\beta + \alpha_i^* + \varepsilon_{it}, \quad i = 1, \ldots, N t = 1, \ldots, T, \tag{1}$$

where α_i^* contains unobserved individual characteristics determining $\log C_{it}$. To correct for possible correlation between α_i^* and the explanatory variables in x_{it}, we follow Mundlak (1978) and Chamberlain (1984) in assuming that:

$$\alpha_i^* = \bar{x}_i\theta + \alpha_i, \tag{2}$$

where $\bar{x}_i$ denotes the time average of x_{it} and where α_i is uncorrelated with x_{it}. Substituting (2) into (1) we obtain our model of interest:

$$\log C_{it} = \beta_0 + x_{it}\beta + \bar{x}_i\theta + \alpha_i + \varepsilon_{it}, \tag{3}$$

where the error term ε_{it} is assumed to be independently identically distributed over individuals and time, independent of all x_{js}. We assume α_i to be normal with expectation zero, variance σ_α^2 and independent of ε_{jt} and $\mathbf{x}_{jt}$ $(\forall i, j, t)$.

Models such as (3) have been analysed in a number of papers trying to model consumer behaviour (see, e.g., MaCurdy, 1981). One way to link this equation to economic theory is to base it on a life cycle model with specific assumptions concerning taste shifters and marginal utility of wealth. Using the framework of MaCurdy (1981), for example, model (3) can be derived making some specific functional form assumptions.

If a complete panel were available the estimation of (3) is straightforward. However, in applied work the panel is likely to be incomplete, i.e. not all relevant variables of each individual are observed in all periods under consideration. Two causes of missing data have to be distinguished. A first reason why observations are missing can be that not all individuals have been asked to report information on all the variables in each period. This will, for example, be the case if the panel is rotating or if the number of individuals included in the panel has been changed during the sample period. A second cause of missing observations is that individuals are not willing (or able) to report on some of the variables.

The data used in this paper are taken from the Expenditure Index Panel conducted by INTOMART, a marketing research agency in the Netherlands. Because many background variables are observed only once a year, we constructed from this monthly panel a panel of yearly observations for April 1984–March 1985, April 1985–March 1986 and April 1986–March 1987. For ease of presentation we refer to these periods as 1984, 1985 and 1986, respectively. More details of the data set can be found in the Appendix. We define the dummy variables a_{it} indicating whether individual i is asked to cooperate in period t (0 if not, 1 otherwise), and r_{it} indicating whether individual i is asked to cooperate in period t (0 if not, 1 otherwise), and r_{it} indicating whether individual i is observed in period t or not.

It is assumed that the decision of the data collecting agency to include an individual in the sample is independent of the disturbances α_i and ε_{it} in (3) but dependence on the exogenous

variables is not excluded. Thus, we do not abstract from the possibility that the data-collecting agency selects individuals from the population on the basis of certain (exogenous) demographic characteristics (e.g. age, education and family composition), for example to obtain representativeness of the sample with respect to these characteristics. Conditional on $a_{it} = 1$, we postulate a response equation representing the decision of an individual to cooperate or not. In particular, we assume that C_{it} is observed if a latent variable r_{it}^* is nonnegative, where r_{it}^* is explained by a (latent) regression equation. The variable r_{it} is generated by:

$$\begin{aligned} r_{it} &= 0 \quad \text{if } a_{it} = 0 \\ r_{it} &= I(r_{it}^* > 0) \quad \text{if } a_{it} = 1 \end{aligned} \tag{4}$$

with

$$r_{it}^* = \gamma_0 + x_{it}\gamma + \bar{x}_i\mu + z_{it}\delta + \xi_i + \eta_{it} \tag{5}$$

where ξ_i is an individual specific effect (independent of x_{it}), and z_{it} contains variables influencing nonresponse but not influencing total consumption, for example the dummy variable $r_{i,t-1}$ which indicates whether one participated in the previous period or not. Note that we consider item nonresponse on C_{it} only, i.e. x_{it} and z_{it} are observed whenever $a_{it} = 1$.

Letting $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$, $\eta_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT})'$ and $\iota = (1, 1, \ldots, 1)'$ of dimension T, we assume that the error terms in (3) and (5) are normally distributed according to:

$$\begin{pmatrix} \iota\alpha_i + \varepsilon_i \\ \iota\xi_i + \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_\varepsilon^2 I + \sigma_\alpha^2 \iota\iota' & \sigma_{\varepsilon\eta} I + \sigma_{\alpha\xi}\iota\iota' \\ & \sigma_\eta^2 I + \sigma_\xi^2 \iota\iota' \end{pmatrix}\right), \tag{6}$$

and this error term vector is independent of x_{jt} $(\forall i, j, t)$. For identification purposes we will normalize $\sigma_\eta^2 + \sigma_\xi^2 = 1$, as usual. The complete model is given by equations (3), (4), (5) and (6). From a statistical point of view, our model is more general than the one considered by Hausman and Wise (1979), who look at the random effects model for two periods with attrition in the second period only. Because their probit equation is derived in a somewhat different way, the covariances between α_i and ξ_i and between ε_{it} and η_{it} are restricted by $\sigma_{\alpha\xi}/\sigma_{\varepsilon\eta} = \sigma_\alpha^2/\sigma_\varepsilon^2$. Our model fits in the theoretical framework presented by Ridder (1990), who deals with the problem of constructing estimation procedures for the general model.

Nonresponse is random and no selectivity bias occurs in consumption function estimates if the unobserved determinants of response are uncorrelated with the unobserved determinants of consumption, i.e. if $\sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$. This will be our null hypothesis, H_0.

3. THE PROBLEM OF NONRESPONSE AND SOME SIMPLE TESTS

In our sample, some of the households do not report total consumption in any of the 3 years under consideration (e.g. because they stay in the panel for less than a year), some report for 1 year, some for 2 years and others for all 3 years. If consumption is not observed for some household in some period this is either due to the fact that a household is not asked (any more) to report its expenditures or to refusal or inability of the household to supply expenditure data (given that it is asked for). Because the data on consumption are collected on a monthly basis it is possible to distinguish between these two types of nonresponse if we are willing to model the way in which the data-collecting agency reacts if a household does not respond. We assume that the household is repeatedly asked to cooperate by the data-collecting agency until it has not been responding for 6 consecutive months. The actual strategy used by INTOMART is more complicated than this, but can probably be closely approximated by this assumption. If

a household is not asked to cooperate in the first month of a year (April, in our case), that household is, by assumption, not asked to cooperate in that year.

Using these assumptions the actual distribution of the households in our sample over these possibilities is given in Table I. Of course a household can belong to a different category in each year. Table II gives some more information on the response patterns in the data set under consideration. We see in this table, for example, that only 113 households are observed in all 3 years, while 129 households are observed in 1984 only. A comparison of Tables I and II shows that no observation on annual consumption for any of the three years under consideration is available for 1024 households. We nevertheless have information on their characteristics from the same data set because they cooperated for at least 1 month. This information will be used to estimate the response process.

According to (5) the response probabilities depend on the variables which determine total consumption and on additional variables z_{it}. In our application, z_{it} contains $r_{i,t-1}$, a dummy variable which is one if the individual participated in the previous period, and $a_{i,t-1}$, a dummy which is one if the individual was asked to participate in the previous period. Thus, we can rewrite (5) as:

$$r_{it}^{*} = \gamma_0 + x_{it}\gamma + \bar{x}_i\mu + r_{i,t-1}\delta_1 + a_{i,t-1}\delta_2 + \xi_i + \eta_{it}. \tag{7}$$

The dummy variables are added to model possible state dependence, according to which the response probabilities of households with the same demographic characteristics and even the same unobserved individual effect ξ_i can differ because, e.g., one household cooperated in the previous period while a second household refused to cooperate. As an illustration, consider three identical households (A, B and C) with the same values for each x_{it} and ξ_i in (7), but

Table I. Characterization of the observations

	1984	1985	1986
1. Observed ($r_{it} = 1$)	307	377	366
2. Not observed; asked ($r_{it} = 0$, $a_{it} = 1$)	204	505	404
3. Not asked ($a_{it} = 0$)	1157	786	898
Total	1668	1668	1668

Table II. Characterization of numbers of households with annual consumption observed at least once

	1984	1985	1986
Complete	113	113	113
Observed in 2 years		115	115
	48	48	
	17		17
Observed in 1 year	129		
		101	
			121
Total observed	307	377	366

with different participation histories. Household A was not asked to cooperate in the previous period (and has $r_{i,t-1} = a_{i,t-1} = 0$). Household B was asked in the previous period, but refused, while household C cooperated. The mere effect of their different participation histories on the response probabilities in the present period is captured by the term $r_{i,t-1}\delta_1 + a_{i,t-1}\delta_2$. The difference in response probabilities between household A and B is caused by β_2, which is expected to be negative since household B refused at an earlier stage, while A was not asked to cooperate before. The difference in response probabilities between household B and C is caused by δ_1, which is expected to be positive, since an earlier refusal is expected to have a negative effect on the response probability in the present period. The difference between household A and C is caused by $\delta_1 + \delta_2$. In case of a heavy response burden, household C may be tempted to leave the panel after one or more waves of participation, in which case there is negative state dependence on response in the previous period ($\delta_1 + \delta_2 < 0$). On the other hand, if household C values the way in which it is forced to keep close track of its expenditures, this may be an incentive to stay in the panel, resulting in a positive state dependence on previous response ($\delta_1 + \delta_2 > 0$). Consequently, we do not have *a priori* expectations concerning the sign of $\delta_1 + \delta_2$.

If households A, B and C only have the same values for each x_{it}, an alternative explanation is possible for the fact that household B (which refused in the previous period) is less likely to be observed in the present period than household C (which cooperated in the previous period). It is not unlikely that the two households differ in unobserved characteristics that persistently affect nonresponse, i.e. they have different values for ξ_i. Using panel data it is possible to disentangle this unobserved heterogeneity from state dependence (see, e.g., Heckman, 1981).

The parameters in our model can be estimated in a wide variety of ways. We assume that the parameters of interest are β and θ in equation (3). If $\sigma_{\varepsilon\eta} = 0$ and $\sigma_{\alpha\xi} = 0$ the efficient estimator of these parameters is the standard random effects (RE) regression estimator (see e.g. Hsiao, 1986, p. 36) on the unbalanced panel. This estimator will not be consistent, however, if $\sigma_{\varepsilon\eta} \neq 0$ or $\sigma_{\alpha\xi} \neq 0$ (see e.g. Hausman and Wise, 1979 or Ridder, 1990). In Verbeek and Nijman (1992), it is shown that the within or fixed effects regression estimator is more robust to nonrandom nonresponse and, in particular, that it will be consistent if $\sigma_{\varepsilon\eta} = 0$ or $x_{it}\gamma + z_{it}\delta$ does not vary over time. This latter situation may be relevant if the variables in (5) that have a non-zero coefficient are time-invariant.

Consistent and efficient estimators of all parameters in the model without imposing the restrictions $\sigma_{\varepsilon\eta} = 0$ and $\sigma_{\alpha\xi} = 0$ can be derived along the lines sketched in Section 7 below, but these estimators are computationally demanding. Therefore, it is very important to have simple procedures which can be used to check the consistency of computationally attractive estimators. Three possibilities are considered in this paper:

1. One can analyse one wave of the panel as a cross-section in order to obtain simple tests of the hypothesis H_0: $\sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$ either using the well-known Heckman (1979) procedure to correct for sample selectivity in cross-sections or using standard ML routines, both of which are readily available in computer packages such as LIMDEP. In Section 4 we will show that the use of these tests requires that there is either no state dependence ($\delta_1 = 0$) or no unobserved heterogeneity ($\sigma_\xi^2 = 0$) in the response process.
2. One can compare the random effects regression estimates on a balanced and unbalanced panel using a Hausman test as suggested in Verbeek and Nijman (1992). If $\sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$ both estimators are consistent and the one based on the unbalanced panel is efficient. Because both estimators are inconsistent under the alternative the power of this test may be limited, however.

3. One can, similar to the Heckman (1979) procedure in the cross-sectional case, add one or more correction terms to the regression equation (3) using an estimated version of the response equation. Under H_0 these correction terms should not be significant. This approach requires numerical integration (over one dimension) in order to compute the correction terms.

We will return to the three possibilities listed above in Sections 4, 5 and 6, respectively.

4. ESTIMATION USING ONE WAVE OF THE PANEL

A simple possibility suggested in the previous section is to start with analysing just one wave of the panel as a cross-section in which lagged variables are also observed. This is the subject of the present section. The advantage of analysing one wave only is that the issue of selectivity for this case has been widely discussed in the literature. The disadvantage is, of course, a loss in efficiency, and moreover the fact that one can no longer distinguish between state dependence and unobserved heterogeneity.

The complete model based on one wave of the panel is given by:

$$\log C_{it} = \beta_0 + x_{it}\beta + \bar{x}_t\theta + \varepsilon_{it}^* \tag{8}$$

$$r_{it}^* = \gamma_0 + x_{it}\gamma + \bar{x}_t\mu + \delta_1 r_{i,t-1} + \delta_2 a_{i,t-1} + \eta_{it}^*, \tag{9}$$

where $\varepsilon_{it}^* = \alpha_i + \varepsilon_{it}$ and $\eta_{it}^* = \xi_i + \eta_{it}$ are normally distributed error terms with mean zero, variances σ^2 and 1, respectively, and correlation coefficient $\rho_{\varepsilon\eta}$. In our application, x_{it} contains the following time-varying variables: the level of education (1–7) of the head of the household, the age and age squared of the head (divided by 100 and 10,000 respectively), the numbers of children between 0 and 5, 6 and 12 and between 13 and 18 and the number of adults in the household.

If $r_{i,t-1}$ were exogenous or if *a priori* $\delta_1 = 0$, equations (8) and (9) constitute the standard sample selection model of Heckman (1976), also known as the Type II Tobit Model of Amemiya (1984). However, if $\sigma_\xi^2 \neq 0$, $r_{i,t-1}$ is correlated with η_{it}^* and standard maximum-likelihood estimators based on one wave of the panel will be inconsistent (cf. Heckman, 1981). Consequently, one has to choose from two alternative approaches to ensure consistency of the ML estimator:

I. Include $r_{i,t-1}$ and assume that there is no unobserved heterogeneity ($\sigma_\xi^2 = 0$).
II. Exclude $r_{i,t-1}$, which is valid in the absence of state dependence ($\delta_1 = 0$).

Note that similar problems do not arise with respect to $a_{i,t-1}$ which is by assumption exogenous.

Standard ways to estimate the sample selection model are the maximum-likelihood method or the two-step estimation procedure put forward by Heckman (1976, 1979), which is computationally more attractive. If the error terms are uncorrelated ($\rho_{\varepsilon\eta} = 0$), the ordinary least-squares estimator for the parameters in (8) is consistent. We first estimated (8) with ordinary least-squares using the 1986 wave of the panel. Then we estimated (8) jointly with (9) with maximum-likelihood both for case I and II. The results[1] show that only few variables have a significant impact on the response behaviour. In particular, these are the number of adults in the household, which has a negative influence on the tendency to respond, and the two dummy variables, $r_{i,t-1}$ and $a_{i,t-1}$, the effects of which confirmed our earlier expectations.

[1] Full estimation results for the 1986 wave are not reported here, but are available from the authors upon request.

Comparison of the least-squares estimates with the ML estimates indicates that—for the analysis of the 1986 wave—selectivity bias is not a serious issue, although the estimates for $\rho_{\varepsilon\eta}$ are $-0{\cdot}59$ $(0{\cdot}16)$ and $-0{\cdot}80$ $(0{\cdot}07)$ for case I and II, respectively (standard errors in parentheses). This is obviously caused by the fact that almost none of the regressors in (8) enters the response equation (9) significantly. Since the point estimates of $\rho_{\varepsilon\eta}$ are highly significant according to standard t-test measures, the null hypothesis of no selectivity bias will be rejected when a Wald test (t-test) is used. However, if we impose the restriction $\rho_{\varepsilon\eta} = 0$ and compare the restricted log-likelihood maximum with the unrestricted one by means of a likelihood ratio test, the test statistic takes the value of $3{\cdot}88$ for case I, which is only slightly significant at a 5 per cent level, and $5{\cdot}92$ for case II. Although the Wald test and the likelihood ratio test are asymptotically equivalent, this is not too surprising. It has been shown in the literature that the numerical value of the Wald test statistic in small samples is highly dependent on the algebraical formulation of the null hypothesis, see e.g. Gregory and Veall (1985), Lafontaine and White (1986) and Phillips and Park (1988). Reformulating the restriction $\rho_{\varepsilon\eta} = 0$ as $\sigma_{\varepsilon\eta} = 0$ reduces the values of the Wald test statistic considerably.

It is important to stress that all estimators in this section are inconsistent if both state dependence and unobserved heterogeneity are present in the response process. With this in mind, one may conclude tentatively that nonresponse, though not completely random with respect to annual consumption, does not seriously distort estimation of the consumption function based on the 1986 wave on the panel. Of course, this does not necessarily imply that inferences based on a simultaneous analysis of all waves of the panel are not distorted by the nonresponse problem.

5. A HAUSMAN TEST ON NONRESPONSE BIAS BASED ON ESTIMATES FROM A BALANCED AND UNBALANCED PANEL

An alternative procedure to test for nonresponse bias is to compare estimates of the consumption function (3) based on the balanced sub-panel of complete observations only, with estimates from the unbalanced panel using a Hausman test. Because both estimators are consistent under H_0 and the one based on the unbalanced panel is efficient, significant differences between the estimates should be caused by a nonrandom response problem if (3) is otherwise correctly specified. An elaborate analysis, including a Monte-Carlo study, of this Hausman test and of related tests for selectivity is given in Verbeek and Nijman (1992).

Since many of the exogenous variables included in our specification do not vary much over time for a given individual, the β parameters are in general not very well identified. Therefore, we will present estimates of $\beta + \theta$ (which is relatively well identified and represents the effect of a permanent change in the explanatory variables on consumption) and θ. Random effects estimates of the parameters $\beta + \theta$ and θ in (3) based on the balanced sub-panel (of 113 households) are presented in the first column of Table III. Although in applied work attention is usually restricted to balanced sub-panels, it is still rather straightforward to analyse unbalanced panels as long as possible selectivity bias is ignored, i.e. as long as $\sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$ is assumed. The random effects estimator for the unbalanced case can easily be obtained from OLS on the model in transformed data, just like in the balanced case, but with transformations that depend on the number of time series observations for each individual (see, e.g., Baltagi, 1985). The estimates for the unbalanced case are presented in the second column of Table III.

Let us concentrate attention on the estimation results from the unbalanced panel first. These estimates, which are consistent under H_0, suggest that households with a head with a higher education consume more than similar households with a low educational level. The relation

Table III. Estimation results consumption function without corrections

	Balanced sub-panel		Unbalanced panel	
$\beta + \theta$				
educ	0·16	(0·02)	0·13	(0·01)
age	10·16	(2·00)	5·57	(0·85)
age-sq.	9·21	(1·92)	−5·09	(0·85)
nkids0–5	0·22	(0·08)	0·14	(0·03)
nkids6–12	−0·08	(0·07)	0·08	(0·02)
nkids13–18	0·18	(0·09)	0·13	(0·03)
nadults	0·35	(0·06)	0·25	(0·02)
θ				
educ	0·15	(0·03)	0·12	(0·02)
age	9·36	(4·18)	−0·84	(2·92)
age-sq.	−8·93	(3·51)	−0·43	(2·45)
nkids0–5	0·25	(0·11)	0·15	(0·06)
nkids6–12	−0·14	(0·10)	0·06	(0·05)
nkids13–18	0·31	(0·12)	0·18	(0·06)
nadults	0·42	(0·09)	0·24	(0·05)
Auxiliary parameters				
intercept	11·04	(0·52)	12·53	(0·20)
$\tilde{\sigma}^2$	0·04	(n.c.)	0·04	(n.c.)
$\tilde{\sigma}_a^2$	0·10	(n.c.)	0·10	(n.c.)
Number of individuals	113		644	
Number of observations	339		1050	

Standard errors in parentheses

between log household consumption and the age of the head of the household appears to be quadratic—all other variables being constant—with a top at the age of 55. Each additional household member has a positive effect on total consumption, the effect being largest for an adult household member and smallest for a child between 6 and 12 years old. The estimate for σ_a^2 of 0·10 is relatively high and implies that 71 per cent of the total error variance can be explained by unobserved individual heterogeneity.

Comparing these results with those based on the balanced sub-panel presented in the first column of Table III, two important points become clear. First, standard errors are substantially higher if the incomplete observations are dropped from the sample. This is obviously due to the information loss. Second, the point estimates in the first and second column differ substantially. For example, the negative sign for the influence of the number of children between 6 and 12 years old is counterintuitive. Consequently, an informal comparison of the estimation results from the balanced sub-panel and the unbalanced panel as performed, e.g., by Björklund (1989) suggests the presence of attrition bias. Note, for example, that the effects of additional household members (both adults and children) are all significantly positive if the unbalanced panel is used. A formal Hausman test for selectivity bias yields the value 30·4 which is significant from a χ^2 distribution with 15 degrees of freedom (the critical value at a 5 per cent level is 25·0 and 30·6 at a 1 per cent level). It is clear from the results in Table III that a Hausman test on a subset of the parameters could reject even more severely.

Of course, the significance of the Hausman test could be due to misspecification instead of attrition bias as well. Therefore, as suggested by Chamberlain (1984), we test the restrictions

(3) imposes on the reduced form parameters π_{sl} in:

$$\log C_{it} = \pi_{0t} + x_{i1}\pi_{1t} + x_{i2}\pi_{2t} + x_{i3}\pi_{3t} + v_{it}. \tag{10}$$

To do this, we first estimate (10) without imposing the error components structure. From these results we estimate the parameters in (3) using a minimum distance technique imposing the following restrictions $\pi_{0t} = \beta_0$, $\pi_{tt} = \beta + \theta/3$ and $\pi_{st} = \theta/3$ $(s \neq t)$, for $t = 1, \ldots, 3$. Using the criterion value it is now straightforward to test these restrictions (see Chamberlain for details). The test statistic takes the value of 57·9 if the balanced panel is used and the value of 25·9 if the unbalanced panel is used, which are both insignificant from a χ^2 distribution with 51 degrees of freedom.

6. TESTS FOR ATTRITION BIAS BASED ON THE ADDITION OF CORRECTION TERMS

In Sections 4 and 5 we tested the hypothesis of no nonresponse bias using standard cross-sectional procedures on a single wave of the panel and using a (quasi) Hausman test on the difference of parameter estimates from a balanced and unbalanced panel. In this section we will consider the third possibility to test for attrition bias referred to in Section 2, viz. the addition of Heckman (1979)-like correction terms to the consumption function in the unbalanced panel. Application of Heckman's two-step estimation method in the panel data case is in principle a straightforward extension of the cross-sectional case.

The central idea here is that, when estimating the model, one is implicitly conditioning upon the outcome of the response process. Since the conditional expectations of the error terms in (3) given $r_i = (r_{i1}, \ldots r_{iT})'$ will be nonzero if $\sigma_{\varepsilon\eta} \neq 0$ or $\sigma_{\alpha\xi} \neq 0$, the estimators may suffer from selection bias. This problem can be solved by including the expectations of the error terms α_i and ε_{it} conditional on the response indicator vector r_i in the model. The remaining error term is—by construction—independent of r_i.

It appears (see Ridder, 1990 or Verbeek and Nijman, 1992) that the conditional expectations of the two components of the error term in (3) can be written as $E\{\alpha_i \mid r_i\} = \sigma_{\alpha\xi} A_{1i}$ and $E\{\varepsilon_{it} \mid r_i\} = \sigma_{\varepsilon\eta} A_{2it}$, with:

$$A_{1i} = \frac{1}{\sigma_\eta^2 + \bar{T}_i \sigma_\xi^2} \sum_{s=1}^{T} a_{is} E\{\xi_i + \eta_{is} \mid r_i\}. \tag{11}$$

and

$$A_{2it} = \frac{1}{\sigma_\eta^2} \left[E\{\xi_i + \eta_{it} \mid r_i\} - \frac{\sigma_\xi^2}{\sigma_\eta^2 + \bar{T}_i \sigma_\xi^2} \sum_{s=1}^{T} a_{is} E\{\xi_i + \eta_{is} \mid r_i\} \right], \tag{12}$$

where $\bar{T}_i = \Sigma_{s=1}^{T} a_{is}$ is the number of periods individual i is asked to cooperate. In the Appendix we present an expression for the conditional expectation $E\{\xi_i + \eta_{it} \mid r_i\}$. The important thing to note is that this conditional expectation is a function of the data and the parameters in the response process only. Consequently, though A_{1i} and A_{2it} are not observed, they can be estimated consistently by replacing the unknown parameters by their estimates obtained from the random effects probit model in (7). The resulting estimated correction terms can be added to (3) and a test for the significance of these terms is a test for nonresponse bias. Note, however, that both maximum-likelihood estimation of the parameters in the response equation and the evaluation of the conditional expectations in the correction terms in (11) and (12) require numerical integration, which makes this two-step procedure much less attractive than in the cross-sectional case.

Because in our application the initial conditions concerning r_{it} are truly exogenous we can set $r_{i0} = 0$ (as well as $a_{i0} = 0$). Consequently, the ML estimator in the probit equation is consistent and asymptotically efficient (cf. Heckman, 1978, 1981). The estimation results for the multivariate probit model based on all three waves of the panel are given in the first column of Table IV. The results show that only few variables have a significant impact on the response probabilities. In particular the age of the head of the household and the number of adults in the household are important determinants of the response probabilities. The response tendency appears to be largest at the age of 59. The unobserved heterogeneity parameter is significant and accounts for 41 per cent of the error variance, but there is little indication of state dependence conditional on participation in the previous period. However, $a_{i,t-1}$ has a significantly negative effect on the response probability in the present period. This implies that households selected by the data-collecting agency tc cooperate in the previous period are less

Table IV. Estimation results response equation and consumption function correcting for selectivity (unbalanced panel 1984–86)

	Probit ML		Feasible GLS	
	$\gamma + \mu$		$\beta + \theta$	
educ	0·02	(0·02)	0·13	(0·01)
age	6·56	(1·84)	5·57	(0·93)
age-sq.	−5·57	(1·88)	−5·10	(0·92)
nkids0–5	−0·03	(0·06)	0·14	(0·03)
nkids6–12	−0·04	(0·05)	0·08	(0·02)
nkids13–18	−0·09	(0·07)	0·13	(0·03)
nadults	−0·19	(0·05)	0·25	(0·02)
	μ		θ	
educ	−0·04	(0·06)	0·13	(0·02)
age	4·36	(9·87)	−0·98	(3·18)
age-sq.	−4·84	(8·38)	−0·36	(2·55)
nkids0–5	0·14	(0·18)	0·15	(0·06)
nkids6–12	−0·03	(0·17)	0·06	(0·05)
nkids13–18	0·05	(0·20)	0·18	(0·06)
nadults	−0·02	(0·14)	0·24	(0·05)
	δ:			
$r_{i,t-1}$	0·22	(0·19)		
$a_{i,t-1}$	−0·44	(0·13)		
Auxiliary parameters				
intercept	−1·32	(0·43)	12·52	(0·25)
$\hat{\sigma}_\xi^2$	0·41	(0·10)		
$\hat{\sigma}_{u\xi}$			0·02	(0·03)
$\hat{\sigma}_{\varepsilon\eta}$			−0·01	(0·06)
Number of individuals	1125		644	
Number of observations	2163		1050	
Log-likelihood	−1391·16			

Standard errors in parentheses (for feasible GLS only valid under $\sigma_{\varepsilon\eta} = \sigma_{u\xi} = 0$).

likely to be responding in this period. This effect is smaller for those who responded in the previous period than for those who did not (as indicated by the positive—though insignificant—coefficient on $r_{i,t-1}$).

Note that if the assumption that the coefficient for $r_{i,t-1}$ is zero (no state dependence) is imposed *a priori*, as in case II of Section 3, the results from the 1986 wave provide consistent estimators of all parameters needed to estimate the correction terms, except for σ_ξ^2 (which cannot be identified from a single cross-section). However, the latter parameter can be estimated through one-dimensional numerical optimization of the likelihood only over σ_ξ^2 with the values of the other parameters replaced by the consistent estimates from Section 3. This is, of course, computationally more attractive than maximum likelihood.

If we substitute the maximum likelihood estimates reported in the first column of Table IV in (11) and (12) the estimated values for A_{1i} and A_{2i} can be used as additional regressors in (3). To obtain consistent estimators with valid standard errors under the null hypothesis ($\sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$) it is most convenient to use a feasible generalized least-squares procedure with an estimate of the variance covariance matrix under the null. Note that the error term in the equation with the correction terms included no longer has an error components structure if the null hypothesis does not hold. The results from the feasible GLS procedure, which are reported in the second column of Table IV, can be used to test for attrition bias using a straightforward F-test or Wald test on the significance of the two correction terms, which yields the insignificant value of 1·11 for the latter. Clearly the assumption of no nonresponse bias is not rejected by the GLS results presented in Table IV. This is also obvious from comparing the results in the second column in Table IV with those in the second column of Table III, obtained without adding the correction terms. The differences are negligible.

7. MAXIMUM-LIKELIHOOD ESTIMATES BASED ON THE COMPLETE MODEL

Efficient estimates of all parameters in the model can be obtained using the maximum-likelihood method. Because numerical integration is required in one or two dimensions for every individual in the sample at each iteration of a high dimensional numerical optimization problem (see Ridder, 1990) this is, though feasible, not computationally attractive. The results for the maximum-likelihood estimator, which required a few dozen iterations from consistent starting values, are given in Table V. These results do not seem to differ very much from the results from the generalized least-squares estimates corrected for potential attrition bias presented in the previous section. Although some of the earlier results, especially those in Section 5, suggested otherwise, the two parameters which model the potential attrition bias, $\sigma_{\varepsilon\eta}$ and $\sigma_{\alpha\xi}$, are both highly insignificant. A likelihood ratio test on the joint significance of $\sigma_{\varepsilon\eta}$ and $\sigma_{\alpha\xi}$ yields the value of 0·46. If we compare the maximum likelihood estimates from the first column of Table V with the random effects estimates without corrections for selectivity bias from Table III the results seem to be fairly similar, which is of course not surprising given the insignificance of the covariances between the two error terms ($\sigma_{\varepsilon\eta}$ and $\sigma_{\alpha\xi}$). Because of the computational burden the ML estimator is not recommended for use in applied work. An efficient estimator which is simpler to compute is the linearized ML or two-step estimator which requires one iteration in the ML procedure from an initial $\sqrt{N}$ consistent estimator only.

According to the results in Table V temporary changes in the explanatory variables have minor effects, because the θ effect is usually dominant over the $\beta + \theta$ effect. For example, average education in the 3 years under consideration has an important effect on consumption, while a change in the level of education from one year to the other has a very small influence.

Table V. Maximum-likelihood estimation results response equation and consumption function correcting for selectivity (unbalanced panel 1984–86)

	Response equation		Consumption function	
	$\gamma + \mu$		$\beta + \theta$	
educ	0·02	(0·02)	0·13	(0·01)
age	6·54	(1·90)	5·37	(0·93)
age-sq.	−5·55	(1·93)	−4·87	(0·93)
nkids0–5	−0·03	(0·07)	0·15	(0·03)
nkids6–12	−0·04	(0·06)	0·09	(0·03)
nkids13–18	−0·09	(0·07)	0·12	(0·04)
nadults	−0·19	(0·05)	0·26	(0·02)
	μ		θ	
educ	−0·04	(0·07)	0·12	(0·02)
age	4·29	(10·95)	−1·28	(2·79)
age-sq.	−4·87	(9·68)	−0·03	(2·52)
nkids0–5	0·14	(0·19)	0·19	(0·05)
nkids6–12	−0·03	(0·17)	0·08	(0·08)
nkids13–18	0·05	(0·20)	0·17	(0·10)
nadults	−0·02	(0·14)	0·26	(0·05)
	δ			
$r_{i,t-1}$	0·23	(0·20)		
$a_{i,t-1}$	−0·45	(0·13)		
Intercepts and error (co)variances				
intercept	−1·31	(0·44)	12·55	(0·25)
$\hat{\sigma}_\xi^2$	0·41	(0·10)		
$\hat{\sigma}_\zeta^2$			0·038	(0·003)
$\hat{\sigma}_u^2$			0·098	(0·007)
$\hat{\sigma}_{u\xi}$			0·008 (0·030)	
$\hat{\sigma}_{\zeta\eta}$			−0·017 (0·063)	
Number of individuals			644	
Number of observations			2163	
Loglikelihood			−1672·94	

Standard errors in parentheses

For the age variables the opposite seems to be the case: the changes in age and age squared are rather important; the levels of age and age squared are not.

If we look at the results in Table V, all variables have the expected signs and the ordering of the effect of the number of children or adults is plausible. Because after all attrition bias is not very important in this application it is not surprising that the final results are close to those obtained using the standard random effects estimator on the unbalanced panel (Table III, column 2) or even using OLS on one wave of the panel only, although the latter estimates are, of course, rather inefficient. However, the estimates which would typically be reported in applied work according to current practice, the random effects estimates based on the balanced panel (Table III, column 1), are not only inefficient compared to the estimates in Tables III (column 2) and V, but moreover have *a priori* implausible signs.

8. CONCLUDING REMARKS

In this paper we analysed the nonresponse bias in estimates of a life cycle consumption model using a Dutch consumer panel. Although it is possible to compute fully efficient estimates of the parameters in the model that we considered, this is computationally demanding. Tests of the importance of the nonresponse problem are fortunately possible using relatively simple procedures. In the current application these tests suggested that there might be an attrition problem, but the evidence was not decisive: the tests based on a single wave of the panel only suggested some attrition bias, but their numerical value depends strongly on the precise way in which the test is carried out. Moreover these tests are invalid if the response mechanism shows both state dependence and unobserved heterogeneity. The Hausman test on the difference between estimates of the expenditure equation of the balanced and unbalanced panel is almost significant at the 1 per cent level. However, addition of Heckman (1979)-type correction terms to the expenditure equation does not show any sign of attrition bias. The final efficient estimates show that there is no nonresponse bias in the present application. As a side-product our results indicate how the efficiency of estimates is affected if one uses either one wave of the panel, a balanced sub-panel, or the unbalanced panel to obtain parameters estimates. Only the standard errors of the latter ones are close to the efficient estimates obtained from the simultaneous maximum likelihood method incorporating possible selectivity of Section 7.

Unfortunately, our analysis does not provide a once and for all clear-cut answer to the question of how nonresponse in panel data should be handled. Based on the experience above, we recommend the use of simple procedures to test for attrition bias before one turns to computationally demanding estimation methods for the general model. Moreover the results show that it is worthwhile to use information from individuals that are not observed in all periods as well, which is not common practice. Of course, this is computationally slightly more demanding than an analysis of the balanced panel, but the extensions are straightforward and do not require numerical integration or other computer time-consuming operations.

APPENDIX

Data

The data used in this paper are taken from the Expenditure Index Panel conducted by INTOMART, a marketing research agency in the Netherlands. In this study we use data of the period April 1984–March 1987. Detailed information about expenditures on different categories of consumer goods is collected on a monthly basis, while data on background variables, such as education, family composition and age are gathered once a year.

The data sets consists of about 800 households per month, of which almost each month a group drops out due to nonresponse. In most months new households are included in the sample, so that the number of observations is approximately constant over the months. Very few households return in the sample if they have not been observed for 1 or more months, i.e. nonresponse leads to attrition in many cases, even though the data-collecting agency repeatedly asks the household to cooperate during periods of absence in the panel. After eliminating some households with unrealistic expenditure patterns, we arrive at a sample of 1668 different households, of which 543 are none of the 3 years asked for their expenditure patterns. Of course, most of these households did participate (at least once) in the yearly survey on background variables. The response pattern of the 1668 households in the sample is given in Table VI. We see, for example, in the table that 48 households responded in 1984, 1985 but

Table VI. Response patterns of households

1984			1985			1986		
1	0	•	1	0	•	1	0	•
113			113			113		
	24	91	115			115		
48			48				48	
17			17			17		
129			129				41	88
	18	83	101				101	
		237			237	102	135	
	5	14		19		19		
	31	48		79			79	
	120	141		261				261
	6				6			6
		543			543			543
307	204	1157	377	505	786	366	404	898

• = Not asked to cooperate ($a_{it} = 0$)
0 = Asked to cooperate, but not willing to ($a_{it} = 1$; $r_{it} = 0$)
1 = Asked to cooperate and willing to ($a_{it} = 1$; $r_{it} = 1$)

refused to respond in 1986, while 83 households were not asked in 1984, responded in 1985 and refused to respond in 1986.

The variables used in the analysis are the following:

1. Log total consumption: the (natural) logarithm of total yearly expenditures in 0·01 Dfl (cents);
2. Education, ranging from 1 (primary education only) to 7 (university degree).
3. Age, age of the head of the household divided by 100;
4. Age-squared (the square of item 3);
5. Nkids0–5, the number of children younger than 6 years;
6. Nkids6–12, the number of children older than 5, younger than 13;
7. Nkids13–18, the number of children older than 12, younger than 19;
8. Nadults, the number of adult household members.

Derivation of the conditional expectations in (11) and (12)

An expression for the conditional expectation of $\xi_i + \eta_{it}$ given r_i can be obtained by using:

$$E\{\xi_i + \eta_{it} \mid r_i\} = \int_{-\infty}^{\infty} [\xi_i + E\{\eta_{it} \mid r_i, \xi_i\}] f(\xi_i \mid r_i) d\xi_i. \tag{13}$$

The conditional expectation in the right-hand side of (13) is simple because the probit error terms are independent conditional upon ξ_i. Thus we obtain:

$$E\{\eta_{it} \mid r_i, \xi_i\} = (2r_{it} - 1)\sigma_2 \frac{\phi\left(\dfrac{B_{it} + \xi_i}{\sigma_\eta}\right)}{\Phi\left((2r_{it} - 1)\dfrac{B_{it} + \xi_i}{\sigma_\eta}\right)} \tag{14}$$

The conditional distribution of ξ_i is given by:

$$f(\xi_i \mid r_i) = \frac{\prod_{s=1}^{T} \Phi\left((2r_{is}-1)\frac{B_{is}+\xi_i}{\sigma_\eta}\right)^{a_{is}} \frac{1}{\sigma_\xi}\phi(\xi_i/\sigma_\xi)}{\int_{-\infty}^{+\infty} \prod_{s=1}^{T} \Phi\left((2r_{is}-1)\frac{B_{is}+\xi}{\sigma_\eta}\right)^{a_{is}} \frac{1}{\sigma_\xi}\phi(\xi/\sigma_\xi)\,\mathrm{d}\xi}, \tag{15}$$

where B_{it} is the deterministic part of the probit equation, i.e.

$$B_{it} = \gamma_0 + x_{it}\gamma + \bar{x}_i\mu + z_{it}\delta. \tag{16}$$

ACKNOWLEDGEMENTS

The authors have benefited from financial support of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO), respectively. Helpful comments by Pim Adang, Gerard van den Berg, Marc Bout, Cheng Hsiao, Anders Klevmarken, Costas Meghir, Geert Ridder and two anonymous referees are gratefully acknowledged.

REFERENCES

Amemiya, T. (1984), 'Tobit models: a survey', *Journal of Econometrics*, **24**, 3–61.

Baltagi, B. H. (1985), 'Pooling cross-sections with unequal time-series lengths', *Economics Letters*, **18**, 133–136.

Björklund, A. (1989), 'Potentials and pitfalls of panel data. The case of job mobility', *European Economic Review*, **22**, 537–645.

Chamberlain, G. (1984), 'Panel data', in Z. Griliches and M. D. Intriligator (eds), *Handbook of Econometrics*, Vol. II, North Holland, Amsterdam, pp. 1247–1318.

Gregory, A., and M. Veall (1985), 'On formulating Wald tests of nonlinear restrictions', *Econometrica*, **53**, 1465–1468.

Hausman, J. A., and D. A. Wise (1979), 'Attrition bias in experimental and panel data: the Gary income maintenance experiment', *Econometrica*, **47**, 455–473.

Heckman, J. J. (1976), 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *Annals of Economic and Social Measurement*, **5**, 475–492.

Heckman, J. J. (1978), 'Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence', *Annales de l'INSEE*, **30/31**, 227–269.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica*, **47**, 153–161.

Heckman, J. J. (1981), 'The incidental parameters problem and the problem of initial conditions in estimating a discrete time-distrete data stochastic process', in C. F. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, pp. 179–195.

Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge University Press, Cambridge.

Lafontaine, F., and K. J. White (1986), 'Obtaining any Wald statistic you want', *Economics Letters*, **21**, 35–40.

MaCurdy, Th. E. (1981), 'An empirical model of labor supply in a life-cycle setting', *Journal of Political Economy*, **89**, 1059–1085.

Mundlak, Y. (1978), 'On the pooling of time series and cross section data', *Econometrica*, **46**, 69–85.

Phillips, P. C. B., and J. Y. Park (1988), 'On the formulation of Wald tests of nonlinear restrictions', *Econometrica*, **56**, 1065–1083.

Ridder, G. (1990), 'Attrition in multi-wave panel data', in J. Hartog, G. Ridder and J. Theeuwes (eds), *Panel Data and Labor Market Studies*, North-Holland, Amsterdam.

Verbeek, M., and Th. E. Nijman (1992). 'Testing for selectivity bias in panel data models', *International Economic Review*, **33**, forthcoming.